CTE Workshop Proceedings, 2026, Vol. **13**, pp. 289–299 10.55056/cte.1438

Generative AI as a historical source: source criticism, citation integrity, and the jagged frontier of digital history

Iryna A. Selyshcheva

Kryvyi Rih State Pedagogical University, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine

Abstract. Between the public release of ChatGPT in late 2022 and 2026, generative artificial intelligence (AI) moved from a computational novelty to a structural feature of historical scholarship, reshaping how primary sources are transcribed, described, analysed, and communicated. This article argues that the decisive methodological shift is not the automation of existing tasks but the arrival of a new kind of object for the historian's craft: the large language model (LLM) itself, which must be read as a historical source rather than trusted as a neutral instrument. Drawing on peer-reviewed evaluations, professional-society guidance, primary legal filings, and documented failure cases, the article develops three connected claims. First, generative models are best understood as an "algorithmic cartography" of the digitised record whose "jagged frontier" of competence maps which pasts have been absorbed into training data and which remain silent. Second, the same architecture that enables transcription of damaged manuscripts and large-scale corpus analysis also produces hallucinations and fabricated citations at rates incompatible with the evidentiary standards of the discipline; recent audits and accountability cases in scholarship, government, and the courts illustrate the stakes. Third, the responsible integration of these tools depends on extending traditional source criticism to the model, on non-negotiable verification of every reference, on cryptographic provenance and Indigenous data-governance frameworks, and on assessment redesign rather than prohibition. The article synthesises evidence across document analysis, public history, and pedagogy to propose a programme for a critically literate, symbiotic historical scholarship.

Keywords: digital history, generative artificial intelligence, large language models, source criticism, hallucination and citation integrity, handwritten text recognition, AI literacy in history education

1. Introduction

The integration of artificial intelligence (AI) into historical research and reconstruction marks a defining methodological shift in the humanities, comparable in reach to the advent of digital history itself. Computational approaches to the past are not new: they extend a tradition of databases, geographic information systems, network analysis, and text mining that matured over several decades. What changed after the public release of large language models (LLMs) to a mass audience in late 2022 is the *kind* of operation a machine can perform on the documentary record. Where earlier digital methods counted, indexed, and visualised words, generative systems now *produce* them, generating text, narrative, structured data, and code in response to natural-language prompts [24, 33]. The digital turn in history has accordingly accelerated, embedding algorithms directly in the everyday workflow of doing history and, with them, a set of epistemological questions that the discipline is only beginning to formalise.

This transition reorients the role of the scholar from the retention and retrieval of knowledge toward the orchestration, supervision, and critical evaluation of machine-generated output. In this new landscape, digital history no longer merely uses machines as digital supports or archival search engines; it must decide how far the products of a probabilistic model can be admitted as evidence, and

0000-0002-4841-6449 (I. A. Selyshcheva)
irina.selischeva2016@gmail.com (I. A. Selyshcheva)
https://kdpu.edu.ua/personal/irinaselischeva.html (I. A. Selyshcheva)

Received	Accepted	Published	Version of record
2026-03-06	2026-03-20	2026-03-21	2026-03-21

under what controls. Generative AI offers unprecedented scale in processing the documentary record, but it simultaneously introduces structural biases, epistemic vulnerabilities, and ethical hazards that threaten the integrity of that record [8, 26]. The professional response has been swift: the American Historical Association (AHA) approved formal *Guiding Principles for Artificial Intelligence in History Education* in 2025 [4, 20] and has taken up AI in teaching and research in its public programming [46], the *Journal of Digital History* devoted an issue to AI and history while *Histories* opened a special issue on AI and historical research [26, 29, 47], and the *American Historical Review* opened a standing call for essays on AI in historical perspective [5].

This article advances a single organising thesis: the most consequential move available to historians is not to treat the LLM as a faster transcription engine but to treat it as a historical source in its own right, subject to the same source criticism the discipline applies to any other artefact. Section 2 develops this argument and the related metaphors of the "jagged frontier" and "algorithmic cartography". Section 3 surveys the empirical evidence on the archival value chain – optical character recognition (OCR), handwritten text recognition (HTR), transcription, and corpus analysis – and foregrounds a documented failure mode that exemplifies why uncritical use is dangerous. Section 4 examines hallucination and citation fabrication and the accountability cases that have followed. Section 5 addresses synthetic media, provenance, and the "liar's dividend". Section 6 turns to bias, colonial silences, and Indigenous data governance. Section 7 considers public history and museums. Section 8 reviews disciplinary governance and pedagogy. Section 9 proposes a programme for symbiotic scholarship.

2. Generative AI as a historical source: the jagged frontier

To evaluate the impact of generative AI on historical practice, scholars require a framework of source criticism designed for the algorithmic age. Rather than treating LLMs solely as neutral utilities, digital historians have begun to conceptualise these systems as complex historical sources. An LLM is a statistical representation of the corpus on which it was trained; it functions, as this article proposes, as an algorithmic cartography of the digitised record – a skewed, predominantly English-language and Global-North subset of collective digital culture rather than a neutral sample of it. That cartography is uneven. Following recent field-experimental work, the competence of contemporary models can be described as a "jagged frontier": prominent peaks of fluent, high-resolution capability stand beside deep troughs of failure, and the boundary between them is neither intuitive nor stable [15]. The metaphor describes variance in *task competence*; crucially, factual distortion is not confined to the troughs, for a model may execute a task expertly while fabricating content within it.

Because these systems generate output by predicting statistically likely continuations rather than by verifying truth, they operate, in the influential formulation of Bender et al. [8], as "stochastic parrots" that excel at plausibility without comprehension. Consequently, when historians query a model about the past, its responses are highly sensitive to the scale, quality, and bias of the underlying training data. This dependence is itself analytically valuable. Tracing where a model succeeds and where it fails maps which histories have been digitised, indexed, and absorbed into global AI pipelines and whose pasts remain obscured. Hutchinson's probe-based study of LLMs as digital tools demonstrates the method concretely, using controlled prompts about figures such as Boudica and Harriet Tubman's Combahee River raid to reconstruct, through the model's behaviour, the contours of its latent historical knowledge [26]. Interrogating the model as an artefact in this way exposes how commercial AI infrastructures can project a false sense of historical certainty over areas of genuine archival silence.

This reframing extends, rather than discards, the methodological floor of digital history. Guldi's account of text mining as a deliberate, theory-laden craft insists that scale and interpretation must be held together, and that the historian's comparative advantage lies precisely in the disciplined evaluation of evidence and disagreement [23, 24]. The same disposition is what source criticism of the model requires: not credulity toward fluent output, and not refusal of a powerful instrument, but

the systematic reconstruction of the conditions under which that output was produced.

3. Re-engineering the archival value chain

The practical applications of generative AI span the entire archival value chain, transforming how primary sources are ingested, transcribed, described, and analysed. Traditional digital methods relied heavily on manual keyword indexing and frequency-based metadata; vision-language models (VLMs) and LLMs now support semantic, stylistic, and multimodal analysis across large, unstructured, and semantically complex corpora. The most acute institutional problem these tools address is the gap between digitisation and description. The pace of digitising archival materials has long outstripped the capacity to generate high-quality descriptive metadata, producing "dark archives" – digitised collections that remain effectively inaccessible because they are unindexed. Generative AI offers a scalable means of automating descriptive cataloguing and metadata extraction, though, as section 4 shows, that automation must be checked rather than trusted.

3.1. Optical character recognition and handwritten text recognition

The strongest empirical evidence for generative AI's archival utility comes from document recognition. In a controlled benchmark on 1,029 pages of eighteenth-century Russian Civil-font books, Levchenko evaluated twelve LLMs against established engines such as Tesseract, Surya, and Transkribus PyLaia, with the best model (Gemini-2.5-Pro) achieving a character error rate of 3.36%, far below the traditional baselines [34]. Comparable results appear for handwritten material: in a study of 1921 Belgian probate records, two-shot prompting of GPT-4o and Claude 3.5 Sonnet produced transcriptions closer to ground truth than EasyOCR, Pytesseract, KerasOCR, and TrOCR [30]. Evaluations of GPT-4 Vision on historical German print likewise report excellent results even where human readers struggle, while documenting practical limitations of runtime, image rotation, and resolution [19]. Even cautious, single-collection studies of non-expert transcription – such as the assessment of GPT-4V on the cover pages of an urban-renewal housing collection – find the approach promising for collections that would otherwise remain untranscribed [31].

Crucially, the same body of work documents a failure mode that should temper enthusiasm. Levchenko [34] reports that one strong general-purpose model, GPT-4o, systematically "over-historicises" its transcriptions, inserting archaic pre-Petrine Slavonic characters – a stylistic anachronism rather than a simple misreading – into eighteenth-century Civil-font texts in roughly 59% of files. The result is output that *looks* more historically authentic while being less faithful to the source. This is a paradigmatic illustration of the jagged frontier: a model can simultaneously exceed traditional engines on aggregate accuracy and introduce confident, period-inappropriate errors that an untrained user would never detect. Automation of provenance-sensitive tasks therefore cannot be separated from expert post-editing.

3.2. Oral history and audiovisual collections

For audio, the picture is similar. Automatic speech recognition based on Whisper has made oral-history collections far more tractable: a comparative evaluation found Whisper the strongest of nine tools tested, with roughly 93% word accuracy on English and about 70% on German interviews [50], and institutional workflows have processed hundreds of hours of audiovisual material for captioning and access [39]. Yet practitioners working with Holocaust testimonies and other oral-history corpora caution that newer model versions can hallucinate more on long, low-quality recordings, so model selection must be matched to the material rather than assumed to improve monotonically with version number [16].

3.3. From recognition to legible corpora

Recognition is only the first step; the deeper transformation is the conversion of labour-intensive collections into machine-readable corpora that support new historical questions. The *On the Books* project at the University of North Carolina at Chapel Hill illustrates the trajectory: by applying OCR and machine learning to more than three million North Carolina session laws, the project produced a plain-text legal corpus and algorithmically identified Jim Crow statutes enacted between 1866/67 and 1967, enabling research that manual methods could not scale to [25]. Such projects show how generative and algorithmic workflows can turn "collections as data" into instruments of historical discovery – while also reminding us, through their careful documentation and human oversight, that scale is valuable only when it is accountable. Cultural-heritage bodies such as Europeana have accordingly published principles that pair enthusiasm for AI-assisted description with explicit attention to training-data bias and attribution [18].

4. Epistemic vulnerabilities: hallucination and citation fabrication

The analytical scale offered by generative AI is inseparable from a set of epistemic hazards rooted in the architecture of the models themselves, which optimise for fluency and coherence rather than factual verification. Generative systems are prone to hallucination – the production of false, misleading, or wholly fabricated content presented with high linguistic confidence. Surveys of the phenomenon distinguish, at the most general level, between intrinsic hallucinations that contradict a provided source and extrinsic hallucinations that cannot be grounded in any source at all [28]. Translated into historical methodology, the same tendencies degrade into recognisable failures: confident assertion of incorrect dates, places, or chronologies; omission of documented events from a summary; outright fabrication of actors or actions absent from the record; the conflation of tentative hypotheses with established fact; and the misattribution of figures or events across divergent periods. None of these is a peripheral software defect; each follows from a system that predicts plausible text without a model of truth.

4.1. Fabricated citations

The most consequential of these failures for scholarship is citation fabrication: the generation of plausible, syntactically flawless references to publications that do not exist. Such fabrications are not formatting slips but holistic inventions shaped by the statistical regularities of academic prose, recombining real author names, real venues, and field-appropriate keywords into a citation that survives visual inspection but collapses under metadata verification. The foundational study by Walters and Wilder quantified the problem: 55% of the bibliographic citations generated by GPT-3.5 were fabricated, falling to 18% for GPT-4, while 43% of the GPT-3.5 citations that did correspond to real works nonetheless contained substantive errors [48]. Larger and more recent audits confirm that the problem persists and varies sharply by model and domain: a large cross-domain preprint analysis of thirteen LLMs across forty research fields reports hallucination rates ranging from roughly 14% to nearly 95% [51].

This contamination has reached peer-reviewed publishing. Recent audits of papers accepted at leading machine-learning venues have identified hallucinated citations in published, reviewed texts: one preprint analysis documents on the order of a hundred fabricated references across dozens of accepted papers at a single conference, and another finds several hundred hallucinated papers across major computational-linguistics meetings [6, 41]. That human review mechanisms failed to catch these references indicates how poorly traditional safeguards are adapted to synthetic reference chains, and how readily a fabricated citation, once published, can propagate through citation graphs and corrupt future evidentiary baselines.

4.2. Accountability in the professions and the courts

The consequences extend well beyond the academy. In a widely reported case, the consulting firm Deloitte agreed to partially refund the Australian Department of Employment and Workplace Relations for a report commissioned under a contract valued at roughly A$440,000, after researchers identified around twenty fabricated references and a fabricated quotation from a federal-court judgment; the revised report disclosed that a generative system (Azure OpenAI) had been used in its preparation, as reported by *Fortune* [38]. In the legal sphere, the English High Court issued a pointed warning in June 2025 after lawyers in two matters cited, respectively, five and eighteen non-existent authorities; the court held that generative tools are not capable of conducting reliable legal research and observed that placing fabricated material before a court could amount to contempt or, in the most serious cases, to perverting the course of justice [17]. For a discipline whose professional standards already hold that "historians do not fabricate evidence" [27], these cases are not distant cautionary tales but a direct statement of the stakes: the verification of every reference is a non-negotiable condition of admitting generative tools into scholarly work.

4.3. Training data and the documentary record

Underlying these failures is a contested question about the provenance of the training corpus itself. The lawsuit filed by The New York Times against Microsoft and OpenAI, which draws on the newspaper's archive extending back to 1851 and was amended to assert that millions of registered works were used in training, has made the opacity of training data a matter of public litigation rather than private inference [22, 32, 45]. For historians, the legal outcome matters less than the methodological lesson: the behaviour of a model is a function of a corpus that is largely undisclosed, evolving, and contested, which is precisely why source criticism of the model is now indispensable.

5. Synthetic media, provenance, and the liar's dividend

Beyond text, generative adversarial networks, diffusion models, and voice-cloning systems have eroded the presumption of authenticity once attached to audiovisual evidence. Convincing historical photographs, video, and audio can now be synthesised at scale with minimal expertise, and the threat is double. The first danger is the direct falsification of the record. The second, more corrosive, is what Chesney and Citron termed the "liar's dividend": as synthetic media become ubiquitous, bad actors can discredit genuine evidence simply by alleging that it is fabricated, so that the mere possibility of a deepfake undermines authentic documentation [12]. For future historians, the risk is an archival environment in which authentic footprints cannot be cleanly distinguished from sophisticated forgeries.

Because forensic detection of manipulated media is fragile – vulnerable to the compression and transcoding artefacts of ordinary platform circulation – durable responses are shifting from *detecting* fakery after the fact toward *constructing* authenticity at the point of capture. The most developed effort is the open standard of the Coalition for Content Provenance and Authenticity (C2PA), which attaches cryptographically signed provenance metadata to digital media, recording origin and edit history in a tamper-evident manifest [14]. Adopting provenance standards of this kind at the point of ingestion offers archives and contemporary-history repositories a way to preserve the evidential value of born-digital records as an active, verifiable property rather than a passive assumption.

6. Bias, colonial silences, and data sovereignty

Because commercial generative systems are trained on vast, internet-scale datasets scraped with little regard for consent or context, they reproduce the biases and silences of that data and can extend the dynamics of dispossession into the digital domain. The pattern is especially visible when models are asked to interpret the colonial past. Analysing more than 3,800 captions that a leading

text-to-image system (Midjourney) generated for one hundred archival photographs of colonial-era "human zoos," Alenichev and colleagues documented a measurable "colonial gaze": the captions recurrently performed cultural erasure (in 54.5% of cases), essentialism (41.6%), othering (28.4%), infantilisation (26.8%), and outright dehumanisation (11.1%), reproducing the perspective of the coloniser rather than the agency or dignity of the people depicted [1]. Such outputs naturalise colonial hierarchies and blunt the historical reality of colonial violence, with direct implications for how marginalised pasts are represented and for contemporary claims to redress. They are, moreover, continuous with the broader documentation that generative models mirror the racial and gendered stereotypes latent in their training corpora [8].

The response from Indigenous scholars and technologists has been to assert governance rather than merely to debug outputs. The CARE Principles for Indigenous Data Governance – Collective benefit, Authority to control, Responsibility, and Ethics – reframe data not as a freely extractable resource but as something over which communities hold legitimate authority, and they are explicitly designed to complement the more familiar FAIR principles (Findable, Accessible, Interoperable, Reusable) of open data [11]. Applied to generative AI, such frameworks imply that the digitisation of cultural heritage must respect relational and localised protocols, and that ingestion of sensitive or sacred materials into public models cannot be treated as a default. Institutional cultural-heritage policy is beginning to reflect this tension: open-access programmes increasingly distinguish between non-commercial reuse and the commercial training of models on collection data [18, 44].

7. Public history, digital humanities, and museums

The transformation of historical practice extends into the public sphere, reshaping how historical knowledge is communicated in museums and at heritage sites. Long stereotyped as backward-looking custodians of static objects, cultural institutions have increasingly adopted AI to turn exhibitions into participatory, data-informed experiences [2, 42]. Some of the most visible interventions treat AI as both medium and subject. Refik Anadol's *Unsupervised*, installed at the Museum of Modern Art from November 2022 to October 2023, used a generative model trained on the museum's own collection to produce continuously evolving imagery, prompting curatorial debate about authorship, originality, and the lineage of training data [7, 43]. Earlier, more service-oriented deployments point in another direction: the Museu do Amanhã in Rio de Janeiro introduced its digital assistant IRIS in 2015 and, in partnership with IBM, the conversational agent IRIS+ in 2017, later adding sign-language translation and audio description to broaden accessibility [37].

These deployments demonstrate genuine gains in engagement and access, but they also import the hazards discussed above. Open-ended conversational agents can expose visitors to historical misinformation, and the simulation of dialogue with historical figures can foster the illusion that the past is fully knowable and unambiguous. Institutional guidance from museum bodies and major collections accordingly emphasises transparency about AI use, human oversight of generated interpretation, and clear policies on the reuse of collection data [2, 10, 44]. The lesson mirrors that of the archive: interactivity is an asset only when the generated content is accountable to the evidentiary and ethical standards of the institution.

8. Disciplinary governance, guidelines, and pedagogy

Because specific, tool-based rules become obsolete within months, professional organisations have converged on high-level ethical frameworks rather than prohibitions. The AHA's *Guiding Principles for Artificial Intelligence in History Education*, approved in 2025, are the most developed disciplinary statement. They were informed by a survey of AHA members in which 68.9% reported having already redesigned at least one course to address generative AI and 92.6% requested formal guidance [4, 20]. Rather than banning the technology, the Principles emphasise historical thinking, AI literacy, and concrete and transparent course policies, including a model table of acceptable and unacceptable

uses; they explicitly reject the submission of AI-generated essays as original work and the use of unverified AI-generated citations [3, 4]. The Royal Historical Society has taken a complementary, curatorial approach, maintaining a regularly updated reading guide on generative AI and history and reporting on classroom pilots [13, 40].

In the classroom, the central challenges are integrity and the limits of detection. Empirical work shows that automated AI-text detectors are unreliable and systematically biased: one widely cited study found that detectors misclassified more than half of essays written by non-native English speakers as AI-generated while rarely flagging native writing, making detection both ineffective and inequitable as a basis for academic-integrity enforcement [35]. The more durable responses redesign assessment so that it is resilient to, or even productively engaged with, generative tools – an argument made early in the discipline, that AI writing tools expose weaknesses in how we set and evaluate work rather than posing a simple threat [21]. Design-thinking and institutional approaches to assessment redesign recommend combining authentic tasks with oral defence and reflective documentation [36, 52], and history-specific pilots demonstrate the pedagogical value of having students critique and revise AI-generated essays as a way of learning both the subject matter and the limits of the technology [9]. Encouragingly, the Royal Historical Society's pilots suggest that history undergraduates adopt generative tools at lower rates than their peers in other fields and raise sophisticated objections to them, noting that the models struggle to construct genuine historical arguments or to ground claims in verifiable evidence [13]. Such scepticism is itself a resource for cultivating critical AI literacy. The discipline's experience with born-digital misinformation reinforces the point: community efforts to identify and clean AI-generated fabrications in reference works show both the scale of the problem and the indispensability of expert human judgement [49].

9. Conclusion: toward a symbiotic historical scholarship

Generative artificial intelligence is neither the end of the historical profession nor a frictionless replacement for human interpretation. It establishes a contested landscape in which computational power and human judgement must enter a deliberate partnership. The evidence assembled here supports a coordinated set of interventions. First, academic programmes should integrate structured instruction in how LLMs are trained, why they hallucinate, and how systemic bias enters their output, so that scholars can read these systems critically as historical sources rather than offloading cognition to them [8, 26]; in the classroom this literacy is most effective when paired with assessment redesigned to be resilient to generative tools – authentic tasks, oral defence, and the critique of AI-generated work – rather than with prohibitions or unreliable AI-text detection [9, 35, 36]. Second, publishers, conferences, and research institutions should implement automated, multi-database reference verification, while retaining manual checking of every citation as a non-negotiable standard – a lesson written in the recent record of scholarly, governmental, and judicial failures [17, 38, 48]. Third, cultural-heritage organisations should re-examine open-access and ingestion policies in light of decolonial data-governance frameworks such as the CARE Principles, so that digitisation does not become a further mechanism of dispossession [11, 44]. Fourth, archives and contemporary-history repositories should adopt cryptographic provenance standards at the point of ingestion to preserve the evidential value of born-digital records against synthetic falsification and the liar's dividend [12, 14].

Underlying all four is the argument with which this article began. The most important shift of the generative era is not that machines can now transcribe a manuscript or summarise a corpus, but that the machine's output depends on a training corpus that is opaque, evolving, and contested – so that the model is itself an object requiring source criticism [26, 29]. Generative systems can accelerate transcription, description, and the structural work of historical research; the interpretive core of the discipline – identifying causal relationships, constructing arguments, exercising historical empathy, and writing meaning into the silences of the record – remains the work of human historians. The future of digital history belongs not to the algorithm alone but to the historian who critically, ethically,

and self-reflexively orchestrates it.

Author contributions

The author confirms sole responsibility for the conception, research, analysis, and writing of this article. The author has read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data availability statement

No new data were created or analysed in this study. Data sharing is not applicable.

Conflicts of interest

The author declares no conflict of interest.

Declaration on Generative AI

During the preparation of this work, the author used large language model–based assistants to support literature search and source triage, the structural organisation and drafting of the manuscript, and the cross-checking of bibliographic references against their primary sources. Every cited source was verified against its original; claims that could not be corroborated were removed. The author reviewed and edited all content and takes full responsibility for the publication's content.

References

[1] Alenichev, A., Shaffer, J.D., Kingori, P., Grietens, K.P., Muldoon, J. and Rocher, L., 2026. 'We can see a savage': a case study of the colonial gaze in generative AI algorithms. *AI & SOCIETY*, 41(4), pp.3413–3435. Available from: https://doi.org/10.1007/s00146-025-02685-0.

[2] American Alliance of Museums, 2024. AI Adolescence. *Museum*. Available from: https://www.aam-us.org/2024/01/16/ai-adolescence-in-museums/.

[3] American Historical Association, 2025. AHA Publishes Guiding Principles for Artificial Intelligence in History Education. AHA News. Available from: https://www.historians.org/news/aha-publishes-guiding-principles-for-artificial-intelligence-in-history-education/.

[4] American Historical Association, Ad Hoc Committee on Artificial Intelligence in History Education, 2025. Guiding Principles for Artificial Intelligence in History Education. American Historical Association. Approved by the AHA Council, 29 July 2025. Available from: https://www.historians.org/resource/guiding-principles-for-artificial-intelligence-in-history-education/.

[5] American Historical Review, 2024. AHR Call for Proposals: AI in Historical Perspectives. AHR History Lab. Rolling call through 30 December 2026. Available from: https://www.historians.org/news-publications/american-historical-review/how-to-submit/ai-in-historical-perspectives/.

[6] Ansari, S., 2026. Compound Deception in Elite Peer Review: A Failure Mode Taxonomy of 100 Fabricated Citations at NeurIPS 2025. Available from: https://doi.org/10.48550/arXiv.2602.05930.

[7] Antonelli, P., Reas, C., Anadol, R. and Kuo, M., 2021. Modern Dream: How Refik Anadol Is Using Machine Learning and NFTs to Interpret MoMA's Collection. Available from: https://www.moma.org/magazine/articles/658.

[8] Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, FAccT '21, p.610–623. Available from: https://doi.org/10.1145/3442188.3445922.

[9] Black, A., 2025. The ChatGPT Exam: Critiquing Generative AI to Assess Learning. *Teaching History: A Journal of Methods*, 49(1), p.34–41. Available from: https://doi.org/10.33043/gg58bfzzg.

[10] B Dikow, R., DiPietro, C., G Trizna, M., BredenbeckCorp, H., G Bursell, M., B Ekwealor, J.T., J Hodel, R.G., Lopez, N., B Mattingly, W.J., Munro, J., M Naples, R., Oubre, C., Robarge, D., Snyder, S., L Spillane, J., Tomerlin, M.J., J Villanueva, L. and E White, A., 2023. Developing responsible AI practices at the Smithsonian Institution. *Research Ideas and Outcomes*, 9, p.e113334. Available from: https://doi.org/10.3897/rio.9.e113334.

[11] Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J. and Hudson, M., 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1), p.43. Available from: https://doi.org/10.5334/dsj-2020-043.

[12] Chesney, R. and Citron, D.K., 2019. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107(6), pp.1753–1820. Available from: https://doi.org/10.15779/Z38RV0D15J.

[13] Clayton, D., Altink, H. and Wilson, E., 2025. Piloting Responsible and Effective Use of Generative AI in Undergraduate History Teaching. Historical Transactions (Royal Historical Society). 16 July 2025. Available from: https://blog.royalhistsoc.org/2025/07/16/piloting-responsible-and-effective-use-of-generative-ai-in-undergraduate-history-teaching/.

[14] Coalition for Content Provenance and Authenticity, 2025. C2PA | Verifying Media Content Sources. Available from: https://c2pa.org/.

[15] Dell'Acqua, F., McFowland, E., Mollick, E., Lifshitz, H., Kellogg, K.C., Rajendran, S., Krayer, L., Candelon, F. and Lakhani, K.R., 2026. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science*, 37(2), pp.403–423. Available from: https://doi.org/10.1287/orsc.2025.21838.

[16] Draxler, C., Heuvel, H. van den, Hessen, A. van, Ircing, P. and Lehečka, J., 2024. Speech Technology Services for Oral History Research. In: I. Anuradha, M. Wynne, F. Frontini and A. Plum, eds. *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*. Torino, Italia: ELRA and ICCL, pp.38–43. Available from: https://aclanthology.org/2024.htres-1.6/.

[17] England and Wales High Court, 2025. Ayinde -v- London Borough of Haringey, and Al-Haroun -v- Qatar National Bank. [2025] EWHC 1383 (Admin), judgment of 6 June 2025 (Dame Victoria Sharp P. and Johnson J.). Available from: https://www.judiciary.uk/judgments/ayinde-v-london-borough-of-haringey-and-al-haroun-v-qatar-national-bank/.

[18] Europeana, 2023. AI in relation to GLAMs Task Force: Report and recommendations. Europeana Pro. Available from: https://pro.europeana.eu/files/Europeana_Professional/Europeana_Network/Europeana_Network_Task_Forces/Final_reports/AI%20in%20relation%20to%20GLAMs%20Task%20Force%20Report.pdf.

[19] Ghiriti, A., Göderle, W. and Kern, R., 2024. Exploring the Capabilities of GPT4-Vision as OCR Engine. In: A. Antonacopoulos, A. Hinze, B. Piwowarski, M. Coustaty, G.M. Di Nunzio, F. Gelati and N. Vanderschantz, eds. *Linking Theory and Practice of Digital Libraries (TPDL 2024)*. Cham: Springer Nature Switzerland, *Lecture Notes in Computer Science*, vol. 15178. Available from: https://doi.org/10.1007/978-3-031-72440-4_1.

[20] Gillis, B., 2025. A Disciplinary Approach to Generative AI in the History Classroom. Perspectives on History (American Historical Association). 24 September 2025. Available from: https://www.historians.org/perspectives-article/a-disciplinary-approach-to-generative-ai-in-the-history-classroom/.

[21] Grigoli, L.R., 2023. Townhouse Notes: Ghosts in the Machine. Perspectives on History (American Historical Association). Available from: https://www.historians.org/perspectives-article/townhouse-notes-ghosts-in-the-machine-march-2023/.

[22] Grynbaum, M.M. and Mac, R., 2023. The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. The New York Times, 27 December 2023. Available from: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

[23] Guldi, J., 2023. *The Dangerous Art of Text Mining: A Methodology for Digital History*. Cambridge: Cambridge University Press. Available from: https://doi.org/10.1017/9781009263016.

[24] Guldi, J., 2024. The Revolution in Text Mining for Historical Analysis is Here. *The American Historical Review*, 129(2), pp.519–543. Available from: https://doi.org/10.1093/ahr/rhae163.

[25] Henley, A., Bruckner, L., Jacobs, H., Jansen, M., Nunez, B., Rodriguez, R. and Wilson, M., 2024. On the Books: Jim Crow and Algorithms of Resistance, a Collections as Data Case Study. *Journal on Computing and Cultural Heritage*, 16(4). Available from: https://doi.org/10.1145/3631128.

[26] Hutchinson, D., 2024. Mapping the Latent Past: Assessing Large Language Models as Digital Tools through Source Criticism. *Journal of Digital History*, 3(1). Available from: https://doi.org/10.1515/JDH-2023-0018.

[27] Jackson, S., 2023. Don't Stop Worrying or Learn to Love AI: A Plea for Caution. Perspectives on History (American Historical Association). 6 November 2023. Available from: https://www.historians.org/perspectives-article/dont-stop-worrying-or-learn-to-love-ai-a-plea-for-caution-november-2023/.

[28] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A. and Fung, P., 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), p.248. Available from: https://doi.org/10.1145/3571730.

[29] Journal of Digital History, 2025. AI & history (Issue n.8). Journal of Digital History (De Gruyter / C^2DH Luxembourg). Available from: https://www.journalofdigitalhistory.org/en/articles.

[30] Kim, S., Baudru, J., Ryckbosch, W., Bersini, H. and Ginis, V., 2025. Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records. Available from: https://doi.org/10.48550/arXiv.2501.11623.

[31] Lee, M. and Hsu, J.H.P., 2024. An Evaluation of GPT-4V for Transcribing the Urban Renewal Hand-Written Collection. *ADHO Digital Humanities Conference 2024 (DH2024), Arlington, Virginia*. Available from: https://doi.org/10.48550/arXiv.2409.09090.

[32] Lee, T.B. and Grimmelmann, J., 2024. Why The New York Times might win its copyright lawsuit against OpenAI. Ars Technica. Available from: https://arstechnica.com/tech-policy/2024/02/why-the-new-york-times-might-win-its-copyright-lawsuit-against-openai/.

[33] Leslie, D., 2025. From Future Shock to the Vico Effect: Generative AI and the Return of History. *Harvard Data Science Review*, (Special Issue 5). Https://hdsr.mitpress.mit.edu/pub/bcp7n3bs.

[34] Levchenko, M.A., 2025. Evaluating LLMs for Historical Document OCR: A Methodological Framework for Digital Humanities. In: I.N. Arachchige, F. Frontini, R. Mitkov and P. Rayson, eds. *Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities*. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, pp.75–85. Available from: https://aclanthology.org/2025.lm4dh-1.7/.

[35] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. and Zou, J., 2023. GPT detectors are biased against non-native English writers. *Patterns*, 4(7), p.100779. Available from: https://doi.org/10.1016/j.patter.2023.100779.

[36] Lye, C.Y., 2025. Towards AI-Resilient Assessment: Applying Design Thinking in Assessment Redesign. SIG-AILTA. Available from: https://sigailta.com/2025/09/09/towards-ai-resilient-assessment-applying-design-thinking-in-assessment-redesign/.

[37] Morena, D., 2018. IRIS+ Part One: Designing + Coding a Museum AI. American Alliance of Museums. Available from: https://www.aam-us.org/2018/06/12/iris-part-one-designing-coding-a-museum-ai/.

[38] Paoli, N., 2025. Deloitte was caught using AI in $290,000 report to help the Australian government crack down on welfare after a researcher flagged hallucina-

tions. Fortune. 7 October 2025. Available from: https://fortune.com/2025/10/07/deloitte-ai-australia-government-report-hallucinations-technology-290000-refund/.

[39] Rao, N. and O'Riordan, S., 2024. *Increasing Accessibility of Audiovisual Content Using Whisper*. (A LYRASIS Catalyst Fund Research Report). LYRASIS. Available from: https://doi.org/10.48609/na33-1y19.

[40] Royal Historical Society, 2025. Generative AI, History and Historians: A Reading Guide. Historical Transactions (Royal Historical Society). Available from: https://blog.royalhistsoc.org/2025/10/02/generative-ai-history-and-historians-a-reading-guide/.

[41] Sakai, Y., Kamigaito, H. and Watanabe, T., 2026. HalluCitation Matters: Revealing the Impact of Hallucinated References with 300 Hallucinated Papers in ACL Conferences. Available from: https://doi.org/10.48550/arXiv.2601.18724.

[42] SEGD-Society for Experiential Graphic Design, 2023. MIT Museum. Available from: https://segd.org/projects/mit-museum/.

[43] Shaffi, S., 2023. 'It's the opposite of art': why illustrators are furious about AI. The Guardian, 23 January 2023. Available from: https://www.theguardian.com/technology/2023/jan/23/ai-generated-art-future-museums.

[44] The Metropolitan Museum of Art, 2026. Open Access at The Met. Available from: https://www.metmuseum.org/about-the-met/policies-and-documents/open-access.

[45] The New York Times Company, 2023. The New York Times Company Plaintiff v. Microsoft Corporation, OpenAI, Inc., OpenAI LP, OpenAI GP LLC, OpenAI LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, and OpenAI Holdings, LLC, Defendants. U.S. District Court, Southern District of New York, No. 1:23-cv-11195-SHS. Complaint filed 27 December 2023; First Amended Complaint 12 August 2024. Available from: https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

[46] Trowbridge, D., 2024. "Historians On": AI in Teaching and Research. AHA Podcast, recorded at the 2024 AHA Annual Meeting, San Francisco. Available from: https://www.historians.org/podcast/historians-on-ai-in-teaching-and-research/.

[47] Valleriani, M. and Gruber, D., 2025. Artificial Intelligence (AI) and Historical Research (Special Issue). Histories. Available from: https://www.mdpi.com/journal/histories/special_issues/S5JI978200.

[48] Walters, W.H. and Wilder, E.I., 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1), p.14045. Available from: https://doi.org/10.1038/s41598-023-41032-5.

[49] Wikipedia, 2026. WikiProject AI Cleanup. Available from: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_AI_Cleanup.

[50] Wollin-Giering, S., Hoffmann, M., Höfting, J. and Ventzke, C., 2024. Automatic Transcription of English and German Qualitative Interviews. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 25(1). Available from: https://doi.org/10.17169/fqs-25.1.4129.

[51] Xu, Z., Qiu, Y., Sun, L., Miao, F., Wu, F., Li, X., Wang, X., Lu, H., Zhang, Z., Hu, Y., Li, J., Jin, L., Zhang, F., Luo, R., Liu, X., Li, Y. and Liu, J., 2026. GhostCite: A Large-Scale Analysis of Citation Validity in the Age of Large Language Models. Available from: https://doi.org/10.48550/arXiv.2602.06718.

[52] Yale Poorvu Center for Teaching and Learning, 2026. AI-Resilient Assessment. Yale University. Available from: https://poorvucenter.yale.edu/teaching/teaching-resource-library/ai-guidance-for-teachers/ai-course-assignment-design/resilient.